DATASCIENCE, LEARNING AND APPLICATIONS

DALAS - Analyse de variance

29 février 2024

Laure Soulier - Nicolas Baskiotis

SCIENCES SORBONNE UNIVERSITÉ

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

MLIA Machine Learning & Deep Learning for Information Access

Analyse de variance

- Cadre général du modèle linéaire
- Variable quantitative à expliquer et une ou plusieurs variables qualitatives explicatives
- Objectif
  - Comparer les moyennes empiriques de la variable quantitatives pour différentes catégories / facteurs (variables qualitatives)
  - Savoir si un facteur ou une combinaison de facteurs (interaction) a un effet sur la variable qualitative

Vocabulaire

- Facteurs : variables qualitatives explicatives
- Niveaux : modalités d'une variable qualitative
- Interaction : combinaison de niveaux

Analyse de variance à un facteur

- Variable quanti X
- Variable Quali M

Exemple : Longueur d'un sandwich selon une boulangerie

	Boulangerie 1	Boulangerie 2	Boulangerie 3
	23,3	18,9	22,5
	24,4	21,1	22,9
	24,6	21,1	23,7
	24,9	22,1	24
	25	22,1	24
	26,2	23,5	24,5
Moy (m_i)	24,75	21,53	23,6

Si on considère chaque échantillon comme issu d'une variable aléatoire X_i suivant la loi des grands nombres de paramètre m_i et σ, le problème est donc de tester :

$$H_0 \quad : \quad m_1 = m_2 = m_3 \tag{1}$$
$$H_1 \quad : \quad \exists i; m_i \neq m_j \tag{2}$$

On peut également poser :

$$x_i^j = m_i + \epsilon_i^j \tag{3}$$

Si H_0 est rejeté, le problème se posera d'estimer m_i.

Trois conditions pour l'ANOVA

- Echantillons indépendants : vérifier les conditions choisies
- Variable quantitative suit une loi normale : test de Shapiro-Wilk (p-value$>$0.05 : normalité)
- Homogénéité de variance dans les niveaux : test de Bartlett (p-value$>$0.05 : homogénéité)

$$\frac{1}{n}\sum_i\sum_j(x_i^j-\bar{m})^2=\frac{1}{n}\sum_i\sum_j(x_i^j-\bar{m}_i)^2+\frac{1}{n}\sum_i\sum_j(\bar{m}_i-\bar{m})^2 \quad (4)$$

Variance totale = moyenne des variances et variance des moyennes :

- Premier terme : variance due au facteur
- Deuxième terme : variance résiduelle

ANOVA test

- Si H_0 est vraie alors la variance due au facteur doit être petite par rapport à la variance résiduelle
- Sinon H_1 est vraie.

$\rightarrow$ Test de Fisher (ratio variance facteur/residuelle)

Exemple des sandwichs

Variation	Variance	ddl	CM	Fobs	p-value
Facteur	31,88	2	15,94	12,31	0,0007
Résiduelle	19,43	15	1,29		
Totale	51,31	17			

p-value < 0.05 : longueurs moyennes sont significativement différentes dans chaque boulangerie.

Extension à deux facteurs

Exemple : Notes des étudiants par matière et par université.

Méthodologie :

- Représenter avec des boxplots (différence des moyennes et variances)
- Faire des graphiques représentant les notes en fonction des matières + code couleur par université (permet de voir les interactions entre facteurs)
- Vérification indépendance, loi normale, homogénéité
- Variance décomposée par 4 termes :

$$V_T = V_R + V_{F1} + V_{F2} + V_{F1\times F2} \tag{5}$$

- Test de Fischer avec p-value sur chaque décomposition de facteur + interaction

ANOVA à deux facteurs

Variation	Variance	ddl	CM	Fobs	p-value
Facteur 1	31,88	2	15,94	12,31	0,0007
Facteur 2	31,88	2	15,94	12,31	0,0007
Facteur 1 et 2	31,88	2	15,94	12,31	0,06
Résiduelle	19,43	15	1,29		
Totale	(somme)				

- Facteur 1 : moyennes significativement différentes
- Facteur 2 : moyennes significativement différentes
- Facteur 1 et 2 : interaction des moyennes non significativement différentes

(valeurs fausses, juste pour comprendre l'idée)

Analyse de covariance

- Variable quantitative à expliquer et variables explicatives à la fois quantitatives (covariables) ou qualitatives (facteurs)
- Objectif
  - Comparer les paramètres des différents modèles de régression estimées pour chaque combinaison
- Même principe sauf que le test de Fisher calcule une p-value sur le modèle de régression et non l'impact des facteurs